TOEIC & TOEFL Vocabulary
Secrets Revealed

JALT 2013 Kobe
Oct 26, 2013
Presenter: Guy Cihi

Lexxica R&D
2-7-8 Shibuya 5F
Shibuya-ku, Tokyo 150-0002
info@lexxica.co.jp
Presentation Outline

1) What is “Coverage?”

2) Corpus Analysis - TOEIC and TOEFL

3) Secrets of TOEIC and TOEFL vocabulary

4) How and why ETS uses esoteric vocabulary

5) How graded readers can best support TOEIC and TOEFL score increases
Coverage

There are specific words that occur most frequently within a particular subject domain.

The most frequently occurring words provide the greatest amount of coverage for a domain.

Focusing on learning missing high frequency words is the fastest way to increase coverage of a domain.
We do our own corpus analysis work

We study exactly which words are required to master each subject area.

- All General English: 13,384 words
- Business English: 8,742 words
- College Entrance: 5,435 words
- High School: 3,552 words
- Elementary: 2,000 basic words
- TOEFL: 7,501 words
- TOEIC: 6,480 words
- IELTS: 5,870 words

Copyright 2013 Lexica
TOEIC Corpus Analysis

1,250,000 total words

14,652 different words

6,480 different words constitute 99% of all occurrences

982 different words constitute 90% of all occurrences.

These 982 are the absolutely essential Super High Frequency words of TOEIC
TOEFL Corpus Analysis

1,250,000 total words

16,736 different words

7,501 different words constitute 99% of all occurrences

1,513 different words constitute 90% of all occurrences.

These 1,513 are the absolutely essential Super High Frequency words of TOEFL
TOEIC and TOEFL are Item Response Theory Proficiency Tests — not English ability diagnostic tests. These tests are not designed to provide meaningful advice for improving English ability.

Students are scored based on their correct responses to questions having known difficulty metrics. The difficulty metrics are established through statistical analysis of all prior uses of each question.
Secret #2

Without a full range of questions from easy to difficult, Education Testing Service “ETS,” would be unable to maintain its bell-curve and generate ‘reliable’ scores.

It is impossible to write statistically difficult questions. Only field testing can identify the difficulty of questions.
95% of test questions are recycled. 5% are new questions that are in the process of being measured for difficulty.

The 95% recycling requirement means that vocabulary on the tests can be accurately predicted.
Secret #4

ETS has never, and likely will never issue a vocabulary guide for any of its major tests including: TOEIC, TOEFL, SAT and GRE.

Why?
Secret #4

Because using difficult words, and irregular definitions, are the best way to create a wide variety of questions at all levels of difficulty.

Publishing an official vocabulary guide would both expose a scoring system vulnerability and defeat the purpose of their tests which is to measure familiarity and proficiency with authentic English.
1/3 of the words in all parts of TOEIC and TOEFL are **not common**, high frequency words in General English.

(¼ of the words in IELTS.)
What kinds of words
# Top 2000 high frequency words of TOEIC and General English

<table>
<thead>
<tr>
<th>TOEIC</th>
<th>General</th>
<th>TOEIC</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability</td>
<td>ability</td>
<td>gain</td>
<td>gain</td>
</tr>
<tr>
<td>able</td>
<td>able</td>
<td>gallery</td>
<td>gall</td>
</tr>
<tr>
<td>aboard</td>
<td>about</td>
<td>gallon</td>
<td>game</td>
</tr>
<tr>
<td>about</td>
<td>above</td>
<td>game</td>
<td>gap</td>
</tr>
<tr>
<td>above</td>
<td>abroad</td>
<td>garage</td>
<td>garage</td>
</tr>
<tr>
<td>abroad</td>
<td>absence</td>
<td>garbage</td>
<td>garden</td>
</tr>
<tr>
<td>absence</td>
<td>absolute</td>
<td>gardener</td>
<td>gas</td>
</tr>
<tr>
<td>absent</td>
<td>absolutely</td>
<td>gasoline</td>
<td>gate</td>
</tr>
<tr>
<td>absolutely</td>
<td>absorb</td>
<td>gaze</td>
<td>gather</td>
</tr>
<tr>
<td>abstract</td>
<td>abuse</td>
<td>gear</td>
<td>gender</td>
</tr>
<tr>
<td>accept</td>
<td>academic</td>
<td>gene</td>
<td>general</td>
</tr>
</tbody>
</table>

- **Blue** indicates words frequent only in the TOEIC corpus.
- **Red** indicates words frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.
What does this mean?

EFL students can’t learn the words they need because they aren’t in their study and reading materials.

(Because study materials are simplified.)
I used to say:

Education Testing Service (ETS) purposefully uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.

(Then I talked to ETS authors and editors)
Now I say:

Education Testing Service (ETS) purposefully uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.
To create new test questions:

Authors are told to search through authentic materials to find texts and dialogs to adapt for the different types of test questions.
To evaluate new test questions:

When finished, the authors and editors do not know how difficult their new questions are.

The only way to find out is for ETS to put them into actual tests alongside questions for which they do know the difficulty.
Testing the test questions:

On every TOEIC and TOEFL test 5% of the questions are new questions that have no affect on scoring.

95% are recycled questions that have known and reliable difficulties that can be used for scoring.
ETS’s primary concern is the **consistency** with which their test scores reflect each respondent’s **relative proficiency** with **authentic** English.
From corpus analysis we confirm:

1/3 of the words on TOEIC and TOEFL tests are low frequency ‘authentic’ vocabulary words.

Vocabulary is the primary reason that one test question is more or less difficult than another.
Note that many of the 1/3 low frequency words have multiple meanings.

<table>
<thead>
<tr>
<th>TOEIC</th>
<th>General</th>
<th>TOEIC</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability</td>
<td>ability</td>
<td>gain</td>
<td>gain</td>
</tr>
<tr>
<td>able</td>
<td>able</td>
<td>gallery</td>
<td>gall</td>
</tr>
<tr>
<td>aboard</td>
<td>about</td>
<td>gallon</td>
<td>game</td>
</tr>
<tr>
<td>about</td>
<td>above</td>
<td>game</td>
<td>gap</td>
</tr>
<tr>
<td>above</td>
<td>abroad</td>
<td>garage</td>
<td>garage</td>
</tr>
<tr>
<td>abroad</td>
<td>absence</td>
<td>garbage</td>
<td>garden</td>
</tr>
<tr>
<td>absence</td>
<td>absolute</td>
<td>gardener</td>
<td>gas</td>
</tr>
<tr>
<td>absent</td>
<td>absolutely</td>
<td>gasoline</td>
<td>gather</td>
</tr>
<tr>
<td>absolutely</td>
<td>absorb</td>
<td>gazee</td>
<td>gear</td>
</tr>
<tr>
<td>abstract</td>
<td>abuse</td>
<td>gear</td>
<td>gene</td>
</tr>
<tr>
<td>accept</td>
<td>academic</td>
<td>general</td>
<td>general</td>
</tr>
</tbody>
</table>

Frequent only in the TOEIC corpus.
Frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.
Typical low frequency definition: **crack**

A line along which something has split without breaking into separate parts: “a crack in the surface.”

An illegal street drug: “possession of crack.”

Very good, esp. at a specified activity: “He’s a crack shot.”

To open something after making a concerted effort: “to crack a safe.”
Typical low frequency definition: crack

A line along which something has split without breaking into separate parts: “a crack in the surface.”

An illegal street drug: “possession of crack.”

Very good, esp. at a specified activity: “He’s a crack shot.”

ETS used this: “…it took several years for Apple to _______ the market.”
A: crack  B: break open  C: secure  D: invert
Why use low frequency definitions?

They are difficult and they are **authentic**.

(ETS doesn’t promise practical English.)
ETS’s advice for scoring higher on TOEIC and TOEFL is to read **authentic texts**.

(Graded readers can’t help because the vocabulary is simplified)
How much authentic text?

Based on incidence of occurrence research by Rob Waring, they’ll need to read 6,250 hours of authentic text in order to meet the lower frequency test words often enough to learn them.
Reading at 70 **authentic** words per minute...

2 hours each day for 8.5 years
Reading at 70 graded words per minute...
Graded readers are general English

- All General English: 18,000 semantemes
- Advanced Graded Readers: 9,000 semantemes
- 99% of Graded Readers: 4,000 semantemes
TOEIC and TOEFL are not general English

All General English
18,000 semantemes

Advanced Graded Readers
9,000 semantemes

99% of Graded Readers
4,000 semantemes

Core

TOEFL
9,000 semantemes

TOEIC
8,000 semantemes

Copyright 2013 Lexxica
TOEIC and TOEFL are not general English

EFL students are here...

99% of Graded Readers
4,000 semantemes

TOEFL
9,000 semantemes

TOEIC
8,000 semantemes
How can graded reading help EFL students prepare for TOEIC and TOEFL?
90% of the words that occur in beginner and intermediate level graded readers are also super high frequency words in the TOEIC and TOEFL domains. Because the tests are timed, students who can process the Super High Frequency words faster enjoy a huge scoring advantage. Graded readers can’t teach vocabulary they don’t contain but, they can help students develop automaticity (instant recognition) for the Super High Frequency words occurring in every TOEIC and TOEFL.
What is the best way to use existing graded readers to improve reading and listening?
Repeated timed aural readings.
Example of a repeated, timed, spoken reading approach. This method is highly effective!

<table>
<thead>
<tr>
<th>Title; Headwords</th>
<th>Minutes</th>
<th>Total words</th>
<th>Words per min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>1</td>
<td>10</td>
<td>62</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>2</td>
<td>9</td>
<td>69</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>3</td>
<td>8</td>
<td>78</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>4</td>
<td>8</td>
<td>78</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>5</td>
<td>8</td>
<td>78</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>6</td>
<td>7</td>
<td>89</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>7</td>
<td>7</td>
<td>89</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>8</td>
<td>6.5</td>
<td>96</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>9</td>
<td>6.5</td>
<td>96</td>
</tr>
<tr>
<td>Good Dog, Bad Dog; 75</td>
<td>10</td>
<td>6.5</td>
<td>96</td>
</tr>
</tbody>
</table>
Implemented properly, a graded speed-reading program can help develop automaticity for the SHF core...

99% of Graded Readers
4,000 semantemes

TOEFL
9,000 semantemes

TOEIC
8,000 semantemes
The WordEngine high speed vocabulary system has been proven to develop automaticity for all of the words...

Core

TOEFL
9,000 semantemes

TOEIC
8,000 semantemes

Learn more than 100 new words per day!
When improved outcomes are important, professionals trust WordEngine to get results!
Average TOEIC score increases

Data Source: Nagoya University of Commerce and Business
University of Niigata Prefecture
Average TOEFL score increases

TOEFL pbt score improvements

Standard School Curriculum: 20
Standard School Curriculum Plus WordEngine: 47

Source: Dr. David Coulson, University of Nigata Prefecture

+135%
Contact Lexxica to start a trial program at your school.

Lexxica R&D
2-7-8 Shibuya 5F
Shibuya-ku, Tokyo 150-0002
info@lexxica.co.jp